Multilingual Disfluency Removal using NMT

Eunah Cho, Jan Niehues, Thanh-Le Ha, Alex Waibel

Institute for Anthropomatics and Robotics,
Karlsruhe Institute of Technology,
Karlsruhe, Germany
firstname.lastname@kit.edu

Abstract

In this paper, we investigate a multilingual approach for speech disfluency removal. A major challenge of this task comes from the costly nature of disfluency annotation. Motivated by the fact that speech disfluencies are commonly observed throughout different languages, we investigate the potential of multilingual disfluency modeling. We suggest that learning a joint representation of the disfluencies in multiple languages can be a promising solution to the data sparsity issue. In this work, we utilize a multilingual neural machine translation system, where a disfluent speech transcript is directly transformed into a cleaned up text.

Disfluency removal experiments on English and German speech transcripts show that multilingual disfluency modeling outperforms the single language systems. In a following experiment, we show that the improvements are also observed in a downstream application using the disfluency-removed transcripts as input.

1. Introduction

The challenges posed by spoken language translation (SLT) have received a great deal of research interest lately. One of these challenges is the handling of speech disfluencies. Unlike written language, spoken language contains disfluencies, such as repetitions, false starts, stutters, etc. Speech disfluencies not only hinder the readability of speech transcripts but also harms the performance of subsequent applications of natural language processing (NLP), including machine translation (MT).

One challenge of speech disfluency modeling is the costly nature of disfluency-annotated data. Most state-of-the-art techniques for disfluency detection rely on supervised data, whose annotation process is expensive. Moreover, disfluency-annotated data is often limited to only certain languages, which brings an even bigger challenge when modeling the speech phenomenon for low resourced languages.

Many of the speech disfluencies, such as stutters, false starts and repetitions, commonly occur in different languages. This research, thus, is motivated by the idea that the data sparsity issue can be remedied if joint representations of speech disfluencies across languages are available. Recently, the characteristics of neural machine translation (NMT) utilizing continuous representations of the input sentences instead of a language-specific, fixed format have brought a lot of attention in building a model for multilingual inputs [1, 2]. The potential benefit from sharing semantic representations of the languages is a strong advantage of multilingual NMT. From this assessment, we expect that multilingual disfluency learning can be a solution to a task with the data sparsity issue. Utilizing the neural network (NN)-based system, we aim to obtain a joint representation of speech disfluencies across different languages.

NMT offers a powerful framework where further speech-cleaning or reconstructing operations are allowed, such as reordering and replacement of words. As an initial work to apply the NMT framework for this task, we establish the effectiveness of this framework on the disfluency removal task.

We use disfluency-annotated English meeting data and German lecture data, and build three separated disfluency removal systems: two single language systems (one trained on the English data, another on the German data) and the multilingual system. In each system, the source side consists of disfluent transcripts. In the multilingual system, it includes German and English disfluent transcripts. Note that the transcripts in each language do not form a parallel or comparative corpus; they are speech transcripts from individual sources. The target side then consists of clean transcripts. Each language, therefore, goes through a monolingual translation process where its disfluencies can be directly removed.

This work will demonstrate that by allowing multilingual learning we can improve the overall disfluency removal performance over the single language learning systems. In order to evaluate the performance, disfluency-removed test sets are compared to manually cleaned transcripts. In addition, we conduct an extrinsic evaluation, where we measure the impact of disfluency removal in a subsequent application. Once a disfluent test set is transformed into a clean one, it will be translated into another language using an MT system. By evaluating the MT performance, we can measure the impact of multilingual disfluency removal in a downstream application.

This paper is organized as follows. In Section 2, a brief overview of past research on disfluency removal and multi-

lingual learning is given. Followed by a brief introduction to speech disfluencies in Section 3, The multilingual disfluency removal scheme is explained in Section 4. Preceding experiments on parameter optimization are also discussed. In Section 5, our experimental setups and the results along with an analysis are given. Finally, Section 6 concludes our discussion.

2. Related Work

Including the early approaches using statistical language model [3], there has been extensive research on the speech disfluency issue. One promising direction was using noisy channel model [4, 5], where it is assumed that fluent text, without any disfluencies, passes through a noisy channel which introduces disfluencies. The noisy channel model is later extended with a tree-adjoining grammar [6]. In this work, the authors used a syntactic parser for building a language model. Sequential tagging has shown a good performance as well. In [7], conditional random fields (CRF) with features from lexical, language model, and parser information have been used for the automatic identification of disfluencies. The authors in [8] compared different modeling techniques for this task, including CRF and maximum entropy (ME) model. Recent efforts have made to evaluate different segmentation and disfluency removal techniques for conversational speech [9], where CRFs have been used extensively as well.

Following a huge success of neural networks in many challenges in NLP, however, there has been a great number of research on disfluency detection using NN. Word embeddings from recurrent neural networks (RNN) have been selected as features in a sequential tagging model for disfluency detection [10]. In [11], the authors investigated the combined approach of CRF and feed-forward neural networks (FFNN) on the disfluency detection and punctuation insertion task. The combination of the two modeling techniques yielded 0.6 to 0.8 BLEU [12] points of improvements over the individual models. One recent study [13] explored RNNs for the disfluency detection and analyzed the effect of context on the performance. The authors showed a good incremental properties of RNNs with low latency. Bidirectional LSTMs have been used for this task in [14], with engineered disfluency pattern match features.

The idea behind multilingual speech processing was also supported by [15]. In [16], for example, the authors showed that the statistical approach for speech processing can actually be ported to other languages with minor parameter tuning. The authors in [2] used encoder-decoder networks for multi-task learning, improving the performance for many-to-many tasks, including machine translation, parsing and imaging captioning. For the translation, they used individual encoders for each language and additional decoders for the multilingual output as well. This scheme, however, does not benefit from the attention-based model. In [1], the authors introduced a one-to-many multilingual NMT system, where one source language can be translated into multiple languages. In this scheme, the attention mechanism is located in the each target language decoder. In [17], the authors introduced an attention-based NMT which can accommodate shared attention mechanism for multilingual translation. While supporting many-to-many translation tasks, they integrated a single attention mechanism into the NMT.

NN for sequence-to-sequence learning [18] and the framework of encoder-decoder networks [19] showed its effectiveness in sequence-to-sequence mapping. The drawback of this approach, difficulty in learning from long sentences, is later addressed by having an attention mechanism [20]. With the attention mechanism, which is now used in many state-of-the-art NMT systems, the system can learn which source tokens to observe more when predicting the next target words. NMT systems achieved a greater performance over the traditional approach of phrase-based machine translation (PBMT) using the same parallel data as observed in recent machine translation campaigns [21, 22].

3. Speech Disfluency

Spoken language differs largely from written language. It often contains self-repairs, stutters, ungrammatical expressions and incomplete sentences or phrases.

Different disfluency categories are thoroughly studied in [23]. A common type of disfluency is the *filler* disfluency, which includes filled pauses, as well as discourse markers. Obvious filler words or sounds would be *uh*, *uhm* or their variants. Representative discourse markers in English are *you know* and *well*, and in German *ja* and *nun*, for example. Edit disfluencies, on the other hand, are often written as **(reparandum) * <editing term> correction.** The reparandum represents the aborted part of the utterance. The interruption point * is the point where the speaker aborts the ongoing utterance and introduces the corrected utterance. Therefore, the reparandum includes repetition, false starts, etc. Before introducing the corrected utterance, often speakers use an editing term, i.e. *sorry*. This scheme has been the bases for disfluency analysis and modeling for many languages [24].

Example sentences from the English and German spontaneous data are shown in Table 1. For the German excerpt, its English gloss translation is also provided. The excerpt shows exact and rough repetitions and filler words, marked with "+/ /+" and <> respectively. We can observe that the speech disfluencies in the two sentences occur in a syntactically comparable structure.

Compared to the aforementioned disfluency categories, the detection of false start has been a more challenging task. Table 2 shows chosen example sentences from the spontaneous data. False starts are marked with "-/ /-". The excerpts in both languages observe a false start, followed by an interruption point, and then the corrected utterance occurred.

It is a challenging task to model disfluencies of this category, since the surface pattern of false start and correction

Table 1: *Example sentences containing filler words and repetitions in English and German.*

English	<uhm> right, +/ they don't /+ <uhm> they don't actually go into ...
German	+/ Es gab /+ <uh> es gibt da drei Prinzipien ... (+/ There was /+ <uh> there is three principles ...)

Table 2: *Example sentences containing false starts in English and German.*

English	-/ what if /- <oh> -/ the /- that's what you want to do?
German	-/ Mit dem recht /- er würde wieder zurückgehen. (-/ With the right /- it would again go back.)

is very sparse and therefore hard to model them without detecting an actual semantic break point. However, the similar patterns of speech disfluencies across different languages suggest a possibility that we could benefit from their gathered information embedded in a common semantic space.

4. Multilingual Disfluency Removal

This work is motivated by the idea that the above mentioned characteristics of spoken language can be observed across different languages. In this work, we aim to develop a multilingual disfluency removal model which can capture the shared aspects of the spontaneous speech in different languages.

4.1. Approach

As an initial work on multilingual disfluency removal, we investigate on the performance boost induced by sharing joint representations of disfluencies across different languages. For this experiment, we used disfluency-annotated English meeting and German lecture corpora.

4.1.1. Data

The English data consists of meetings and on-line lectures as described in [25]. The German data includes university lectures as introduced in [26]. Speech disfluencies in each corpus are annotated by human-annotators, following the same guidelines. The meeting corpus, however, includes an additional disfluency class `interruption`, due to the nature of multi-party conversation. In this work, we merged the disfluency class `interruption` and the `non-copy` class, which covers false starts mainly. The detailed description of such classes are further given in [25]. The cleaned-up data set is prepared by removing disfluencies according to the manual annotation. Detailed statistics of the data is shown in Table 3.

Using the NMT framework, the source side represents disfluent transcripts, while target side the cleaned-up transcripts in the same language, therefore including less tokens. Note that the tokens here denote words and punctuation marks, before any operation for generating sub-words is applied.

Table 3: *Data statistics in number of tokens.*

	Train		Test	
	Disfluent	Clean	Disfluent	Clean
English	97,547	85,761	17,046	13,818
German	97,833	85,955	29,510	25,665

As shown in the table, the selected conversational speech transcripts contains a great deal of disfluencies, reaching up to around 12% of disfluency rate for each training data. The English test data consists of multi-party meeting data purely, which is extremely disfluent. The training data, which also includes web-based lecture data, therefore observes a lower disfluency rate than the test data.

Inspired by [5], the disfluency removal task in this work is considered as a translation process, where the disfluent speech is translated into a clean speech. In the parallel data for the NMT training, source side contains disfluent English/German speech transcripts. Each line of the target side is the disfluency cleaned-up version of the source sentence, whose disfluencies are annotated by human annotators.

A major issue in speech disfluency modeling is data sparsity. Thus, an investigation on efficient usage of available data is necessary. In this work, we applied a word-splitting algorithm to the extent where we can represent uncommon words in a letter based form essentially, and only very common words unsplit. This promotes sharing parameters as much as possible. Detailed investigation on this will be shown in the following Section.

4.1.2. Neural Machine Translation Setup

All disfluency removal experiments are conducted using the NMT framework `nematus`[1], which is an attention-based encoder-decoder model for NMT. We generated the sub-word units using byte-pair encoding (BPE), as described in [27]. In order to have a smallest possible unit mostly for less common words and allow common words be modified or added, we set the number of BPE merging operations low at 150. Detailed analysis on this issue will be given in Section

[1] `https://github.com/rsennrich/nematus`

4.2. Our preceding experiments showed that the disfluency model benefits from language independent sub-word embeddings. Therefore, each sub-word will share representations across different languages without a language specific representation. The detailed description on the preceding experiments will be given in Section 4.2.1.

When training all NMT systems, we accepted almost all sentence pairs except for extraordinarily long sentences. The sentences in different languages are then shuffled inside every minibatch. All NMT networks have the empirically chosen 60-dimensional embeddings for each token with dropout at every layer with the probability of 0.2 at the embedding and hidden layers and 0.1 at the disfluent and clean output layers. The systems are trained using gradient descent optimization with Adadelta [28] on minibatches of size 80.

4.2. Multilingual Learning

In order to share as many parameters as possible, we applied a low number setting in BPE on a joined corpus of English and German, so that few common words are treated independently. Most letters are treated similarly.

When using the chosen BPE operation setting, for example, unsplit words are usually frequently used articles, prepositions, personal pronouns, and conjunctions i.e. "the", "and", and "you" in English, "die", "und", and "wir" in German. Using such extensive word-splitting mechanism, the English example sentence given in Table 2 will be written as follows.

w+hat i+f o+h you the th+at+'s w+hat you w+an+t to d+o ?

The sign "+" denotes a split point within a word, which is followed by a white space during the training and testing. While most of less common words are split into smaller units to share the parameters across languages, we also observed that having an identical sequence longer than two split units between two languages was highly unlikely.

From the given example above for example, the sequence of "␣i+ f␣" occurs only once in the German data, as an actual English word. The sequence of "␣w+ an␣" occurs only a few times, as a part of a verb *wandeln (En. gloss: to stroll)* and its conjugatives.

Using the empirically chosen splits, 42.4% of tokens occurring in the training data are left unsplit. This encourages the generation process easily separated for each language.

4.2.1. Parameter Optimization

In this section, we analyze performance of the preceding experiments. We evaluated different sub-word splitting sizes as well as word embedding sizes on the validation data prior to applying one to the test data. In order to show that the approach of splitting less common words extensively while leaving very frequent words intact performs better, for example, we built a character-based multilingual disfluency removal system. Different degrees of sub-word operations have been tried out, and the effectiveness of the language specific representation is investigated as well.

All preceding experiments are evaluated in BLEU on the validation set. Thus, once the disfluent validation set is *translated* into the cleaned transcript, the performance is evaluated compared to the transcript whose disfluencies are all removed according to the human annotation. Validation set consists of 1,400 sentences from the training data, 700 sentences for each language.

The first investigation was on the sub-word size. For this, we changed the sub-word operation settings in the multilingual disfluency removal system and compared the validation data scores. Table 4 shows the results.

Table 4: *Number of sub-word tokens in the training data and the disfluency removal performance.*

System	No. tokens	Dev
Sys 1	971K	78.97
Sys 2	498K	92.59
Sys 3	465K	92.01
Sys 4	372K	92.37

The first system is the character-based system, where each character is handled as a separate token. We can see that representing all words on characters does not yield a good performance. Applying a splitting operation where we have 498K of tokens in the training data, we achieved a better performance on the validation set. Decreasing the number of sub-word tokens in the training data, however, does not result in a great difference between systems. Thus, we chose the split operation setting of System 2 for our all experiments.

The second investigation was on the language specific representation. We analyzed whether supporting a shared sub-word representation in multiple languages is more helpful or it would yield a better performance to have a language specific representation for each sub-word. For this, we applied a language specific ID (either *de_* or *en_*) for each BPE-applied sub-word. Therefore, the example sentences from Table 2 would be written into *en_w+ en_hat en_i+ en_f* ... for English, and *de_m+ de_it de_d+ de_em* ... for German.

Table 5: *Investigation on the effectiveness of a language specific representation.*

System	Dev
Sys 2	92.59
+ LangID	91.05
Sys 3	92.01
+ LangID	91.75

In our experiments it was shown that the model is benefiting from language independent sub-word embeddings, as shown in Table 5. Sharing this information across different

languages seems to be more helpful for the disfluency removal performance.

In this paper, we used the configuration of System 2 for our multilingual disfluency learning setup. Thus, the vocabulary is shared between the two languages by applying the sub-word operation on the joint corpus of German and English. In the multilingual system, a single system is used for disfluency removal for both languages, using the shared vocabulary. In the next section, we will compare the performance of this multilingual system against the single language disfluency removal systems.

5. Experiments and Results

In this section, we describe the experimental setups of the multilingual disfluency removal model.

5.1. Evaluation Setup

Before training, preprocessing is applied for both languages, including smart-casing and normalization. Details of preprocessing can be found in [29]. In this work, we build two single language disfluency removal systems (one for each language) and a multilingual disfluency removal system.

Each disfluent test set is transformed into clean data, in the same language, by using the disfluency removal NMT as described in Section 4.1.2. We apply both intrinsic and extrinsic evaluations, so that we can see the performance of disfluency removal itself as well as can measure its impact in the overall translation process.

The performance of the disfluency removal is evaluated using BLEU [12] by comparing it to the cleaned-up manual transcript. While deploying NMT for disfluency removal task promotes more fluent and natural output in the target language directly, it does not guarantee to generate the output sentence using the same vocabularies as the source sentence, which introduces a challenge in the evaluation in terms of binary classification. Therefore, in this work, we aim to evaluate the similarity between the generated hypothesis and the reference using BLEU.

The disfluency-removed test data is then translated into another language. German test data is translated into English, while English test data is translated into French. The PBMT systems used for translation between different languages are described below.

5.1.1. English→French MT

The system is built using parallel data from EPPS, NC, and TED talks[2]. The parallel data sums up to 2.3M sentences. Among two word-based language models (LMs), one LM is trained on only TED data in order to adapt the domain into spoken language translation. Another LM is trained on the French side of all parallel data. In addition to a bilingual LM [30], an additional LM trained on part-of-speech (POS) is used. The tags are generated by TreeTagger [31]. A short-range reordering is modeled using the POS-based reordering [32]. The system is optimized on an extra TED data using [33]. This baseline system is described in detail in [34].

[2] http://www.ted.com

5.1.2. German→English MT

Translation models are trained on the parallel data from the WMT 2016 evaluation, including the cleaned-up crawl data. The parallel data sums up to 3.8M sentences. Three word-based LMs are deployed in this system; a word-based 4 gram LM, a bilingual LM, and another LM built on a selected data based on cross entropy as described in [34]. In addition to this, we used two word class-based LMs. For this, we use 100 and 1,000 automatically generated clusters for each, as shown in [35]. Different word order between German and English is modeled using the long-range POS-based pre-reordering [36] and lexicalized reordering. The parameters are optimized on the test2014 of the WMT evaluation campaign. The detailed description of the baseline system can be found in [29].

5.2. Results

Table 6 shows the performance of multilingual NMT for disfluency removal. As described in Section 5.1, the performance is measured in BLEU.

Table 6: *Disfluency removal performance.*

System	English	German
Baseline	74.37	78.03
+ no *uh*	76.82	84.90
Single language NMT sys.	81.56	89.61
Multilingual NMT sys.	**83.57**	**90.75**
CRF-based single language sys.	78.78	-

In the baseline setup, we evaluate the test data with its all disfluencies kept, against the cleaned-up manual transcript. We provide another baseline, where simply all trivial disfluencies, i.e. *uh* and *uhm*, are removed. Using the language-dependent NMT for disfluency removal systems, we can see that the test data achieves improved similarity to the manual transcript. Finally, when using the multilingual NMT for the task, we achieve the best performance of both of the languages.

We compare the performance to the CRF-based approach using the identical English meeting data, following the work in [25]. Even though the CRF-based model showed improvements compared to the baseline systems, both NMT-based systems largely outperformed the sequential tagging model. This improvement indicates the effectiveness of the NMT framework for this task.

In order to measure the impact of disfluency removal in a subsequent application, we translated the disfluency-removed test data into a different language and evaluated the

performance in BLEU. Table 7 shows the results.

Table 7: *Impact of disfluency removal in machine translation.*

System	English	German
Baseline	17.08	21.58
+ no *uh*	17.75	23.46
Single language NMT sys.	19.36	24.34
Multilingual NMT sys.	**19.59**	**24.43**
CRF-based single language sys.	18.22	-
Oracle	21.38	25.22

As before, we provide two baseline scores. Simply removing trivial filler words already brings a big improvement over the first baseline. The multilingual disfluency removal system brings 1 to 1.8 BLEU points of improvement over the second baseline, outperforming the improvements achieved when using the single language disfluency removal NMTs. An additional comparison was made to measure the CRF-based single language system's impact on the following MT performance. Consistent to the intrinsic evaluation performance, the NMT-based disfluency removal systems largely outperforms the CRF-based system. Especially the multilingual system brought about 1.4 BLEU points of improvement, compared to the sequential tagging model.

The last row of the table shows oracle scores of the disfluency removal task. For this experiment, we removed all disfluencies in English and German spontaneous speech according to the human annotation. The oracle sets are therefore identical to the reference used in Table 6. For both language directions, we show that multilingual disfluency removal system reaches the closest performance to the upper bound of the experiment.

6. Conclusion

In this paper, we presented an initial approach to model multilingual disfluency removal. Since one of the biggest challenges in disfluency modeling is the sparsity of annotated data, we propose to use disfluency-annotated data from different languages to their shared semantic representations. Inspired by recent success, we used a multilingual neural MT framework in order to promote a joint representation of disfluencies across different languages.

The experiments on English and German spontaneous speech transcripts showed that allowing multilingual training data for disfluency removal indeed improves the removal performance. In addition, the multilingual disfluency removal system outperformed the single language systems in the extrinsic evaluation for a subsequent application as well.

As an initial work of using NMT framework for multilingual disfluency removal task, we believe that this research can be extended into further tasks of spoken language processing, e.g. sentence reconstruction, reformulation, or stylistic change of the text.

7. Acknowledgements

This work was supported by the Carl-Zeiss-Stiftung. The research by Thanh-Le Ha was supported by Ministry of Science, Research and the Arts Baden-Württemberg.

8. References

[1] D. Dong, H. Wu, W. He, D. Yu, and H. Wang, "Multi-task learning for multiple language translation," in *Proceedings of the 53rd Annual Meeting of the ACL and the 7th International Joint Conference on Natural Language Processing*, Beijing, China, 2015, pp. 1723–1732.

[2] M.-T. Luong, Q. V. Le, I. Sutskever, O. Vinyals, and L. Kaiser, "Multi-task sequence to sequence learning," in *International Conference on Learning Representations (ICLR)*, San Juan, Puerto Rico, 2016.

[3] P. A. Heeman and J. F. Allen, "Speech repairs, intonational phrases, and discourse markers: Modeling speakers' utterances in spoken dialogue," *Computational Linguistics*, vol. 25, no. 4, pp. 527–571, 1999.

[4] M. Honal and T. Schultz, "Correction of Disfluencies in Spontaneous Speech using a Noisy-Channel Approach," in *Proceedings of the European Conference on Speech Communication and Technology (EUROSPEECH 2003)*, Geneva, Switzerlands, 2003.

[5] S. Maskey, B. Zhou, and Y. Gao, "A Phrase-Level Machine Translation Approach for Disfluency Detection using Weighted Finite State Tranducers," in *Proceedings of the Ninth International Conference on Spoken Language Processing (INTERSPEECH 2006-ICSLP)*, Pittsburgh, Pennsylvania, USA, 2006.

[6] M. Johnson and E. Charniak, "A TAG-based Noisy Channel Model of Speech Repairs," in *Proceedings of the 42nd Meeting of the Association for Computational Linguistics (ACL 2004)*, Barcelona, Spain, 2004.

[7] E. C. Fitzgerald, "Reconstructing spontaneous speech," Ph.D. dissertation, Johns Hopkins University, Baltimore, Maryland, USA, 2009.

[8] Y. Liu, E. Shriberg, A. Stolcke, D. Hillard, M. Ostendorf, and M. Harper, "Enriching speech recognition with automatic detection of sentence boundaries and disfluencies," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 5, pp. 1526–1540, 2006.

[9] H. Hassan, L. Schwartz, D. Hakkani-Tür, and G. Tur, "Segmentation and Disfluency Removal for Conversational Speech Translation," in *Proceedings of the 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, Singapore, 2014.

[10] E. Cho, T.-L. Ha, and A. Waibel, "CRF-based Disfluency Detection using Semantic Features for German to English Spoken Language Translation," in *Proceedings of the International Workshop for Spoken Language Translation (IWSLT 2013)*, Heidelberg, Germany, 2013.

[11] E. Cho, K. Kilgour, J. Niehues, and A. Waibel, "Combination of nn and crf models for joint detection of punctuation and disfluencies," in *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Dresden, Germany, 2015.

[12] K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: A method for automatic evaluation of machine translation," in *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL 2002)*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, 2002, pp. 311–318.

[13] J. Hough and D. Schlangen, "Recurrent neural networks for incremental disfluency detection," in *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH 2015)*, Dresden, Germany, 2015.

[14] V. Zayats, M. Ostendorf, and H. Hajishirzi, "Disfluency detection using a bidirectional lstm," in *Proceedings of the Sixteenth Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, San Francisco, CA, USA, 2016.

[15] P. Fung and T. Schultz, "Multilingual spoken language processing," *IEEE Signal Processing Magazine*, vol. 25, no. 3, pp. 89–97, 2008.

[16] M. Honal and T. Schultz, "Automatic disfluency removal on recognized spontaneous speech-rapid adaptation to speaker dependent disfluencies." in *ICASSP*, 2005.

[17] O. Firat, K. Cho, and Y. Bengio, "Multi-way, multilingual neural macinspiredhine translation with a shared attention mechanism," in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2016)*, San Diego, CA, USA, 2016.

[18] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in neural information processing systems*, 2014, pp. 3104–3112.

[19] K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, Doha, Qatar, 2014.

[20] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2015.

[21] M. Cettolo, J. Niehues, S. Stüker, L. Bentivogli, and M. Frederico, "Report on the 12th IWSLT evaluation campaign," in *Proceedings of the Eleventh International Workshop on Spoken Language Translation (IWSLT 2015)*, Da Nang, Vietnam, 2015.

[22] O. Bojar, R. Chatterjee, C. Federmann, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, V. Logacheva, C. Monz, *et al.*, "Findings of the 2016 conference on machine translation (wmt16)," in *Proceedings of the First Conference on Machine Translation (WMT 2016)*, Berlin, Germany, 2016.

[23] E. E. Shriberg, "Preliminaries to a theory of speech disfluencies," Ph.D. dissertation, University of California at Berkeley, Berkeley, California, USA, 1994.

[24] W. Wang, A. Stolcke, J. Yuan, and M. Liberman, "A cross-language study on automatic speech disfluency detection." in *HLT-NAACL*, 2013, pp. 703–708.

[25] E. Cho, J. Niehues, and A. Waibel, "Machine Translation of Multi-party Meetings: Segmentation and Disfluency Removal Strategies," in *Proceedings of the International Workshop on Spoken Language Translation (IWSLT 2014)*, Lake Tahoe, California, USA, 2014.

[26] E. Cho, S. Fünfer, S. Stüker, and A. Waibel, "A Corpus of Spontaneous Speech in Lectures: The KIT Lecture Corpus for Spoken Language Processing and Translation," in *Proceedings of the 9th edition of the Language Resources and Evaluation Conference (LREC 2014)*, Reykjavik, Iceland, 2014.

[27] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, Berlin, Germany, 2015.

[28] M. D. Zeiler, "Adadelta: an adaptive learning rate method," in *CoRR*, 2012.

[29] T.-L. Ha, E. Cho, J. Niehues, M. Mediani, M. Sperber, A. Allauzen, and A. Waibel, "The karlsruhe institute of technology systems for the news translation task in wmt 2016," in *PRoceedings of the First Conference on Machine Translation (WMT 2016)*, Berlin, Germany, 2016.

[30] J. Niehues, T. Herrmann, S. Vogel, and A. Waibel, "Wider Context by Using Bilingual Language Models in Machine Translation," in *Proceedings of the Sixth Workshop on Statistical Machine Translation (WMT 2011)*, Edinburgh, Scotland, 2011.

[31] H. Schmid, “Probabilistic Part-of-Speech Tagging Using Decision Trees,” in *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, England, 1994.

[32] K. Rottmann and S. Vogel, “Word Reordering in Statistical Machine Translation with a POS-Based Distortion Model,” in *Proceedings of the International Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skövde, Sweden, 2007.

[33] F. J. Och, “Minimum Error Rate Training in Statistical Machine Translation,” in *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003)*, Sapporo, Japan, 2003.

[34] T.-L. Ha, J. Niehues, T. Herrmann, M. Mediani, E. Cho, Y. Zhang, I. Slawik, and A. Waibel, “The KIT Translation Systems for IWSLT 2013,” in *Proceedings of the International Workshop on Spoken Language Translation*, ser. IWSLT 2013, Heidelberg, Germany, 2013.

[35] F. J. Och, “An efficient method for determining bilingual word classes,” in *Proceedings of the Ninth Conference on European chapter of the Association for Computational Linguistics (EACL 1999)*. Bergen, Norway: Association for Computational Linguistics, 1999, pp. 71–76.

[36] J. Niehues and M. Kolss, “A POS-Based Model for Long-Range Reorderings in SMT,” in *Proceedings of the Workshop on Statistical Machine Translation*, ser. WMT 2009, Athens, Greece, 2009.